Platform for Advanced Scientific Computing

Università della Svizzera italiana — InterDisciplinary Institute of Data Science IDIDS

Università della Svizzera italiana — Facoltà di scienze economiche — Social Network Analysis Research Center SoNAR-C

Università della Svizzera italiana — Faculty of Informatics — Institute of Computational Science ICS

Scalable MCMC algorithm for the accurate estimation of Exponential Random Graph Models

Byshkin Maksym[1], Alex Stivala[2], Antonietta Mira[1], Garry Robins[2], Alessandro Lomi[1]

[1)]Università della Svizzera italiana, Switzerland
[2)]The University of Melbourne, Australia

The increase in the number and size of large networks requires novel efficient methods for their analysis. Exponential Random Graph Models (ERGMs) are exponential family of probability distributions for network structure. The empirical scope of ERGMs is limited by the fact that Maximum Likelihood Estimation (MLE) of model parameters can be obtained only for relatively small networks with a few thousand nodes at most. We propose a new MCMC approach for maximum likelihood estimation of model parameters. In contrast to existing estimation approaches (Bayesian, MCMCMLE and Method of Moments) the approach we propose does not rely on a large number of MCMC simulations to compute equilibrium network configurations. Rather, the Equilibrium Expectations (EE) approach that we propose is based on properties of equilibrium distributions of Markov chains. We implement the EE approach by designing a simple and very efficient algorithm for MLE of ERGM parameters. The suggested estimator is first tested on simulated networks. We compare the results of estimation with results obtained by the Method of Moments and show that the accuracy of estimates obtained via EE is comparable to estimates obtained with the Method of Moments. We demonstrate that the increased speed of the EE algorithm relative to existing methods allows a dramatic increase in the size of networks that can have ERGM parameters estimated by Maximum Likelihood. The empirical value of the EE algorithm is demonstrated also in a study of large biological and social networks.

Maximum Likelihood parameter estimation

$$E_\pi\left(z_A(x)\right) = z_A(x_{obs})$$

x_{obs} is the observed data

$z_A(x)$ is a set of A functions from the space of states x that measure the properties of the model that are theoretically or empirically relevant (statistics)

$E_\Theta(z_A(x))$ are expected statistics that may be obtained by Metropolis-Hastings algorithm if the simulation time is larger than the burn-in time

$$E_\Theta(z_A(x)) = \sum_x z_A(x)\pi(x,\Theta)$$

$\pi(x)$ is the probability of state x

$$\pi(x) = \frac{1}{k}\exp\left(\sum_A \Theta_A z_A(x)\right)$$

Θ_A are model parameters to be estimated

k is a normalizing constant that cannot be computed directly due to complexity

$$k = \sum_x \exp\left(\sum_A \Theta_A z_A(x)\right)$$

Metropolis-Hastings algorithm

$q(x \to x')$ proposal

$$\alpha(x \to x', \Theta) = \min\left\{1, \frac{q(x' \to x)\alpha(x', \Theta)}{q(x \to x')\alpha(x, \Theta)}\right\}$$ acceptance probability

$$P(x \to x', \Theta_A) = q(x \to x')\alpha(x \to x', \Theta_A)$$ transition probability

Basic sampler

Proposal:
A network dyad is chosen uniformly at random and its value is flipped

$q(x \to x') = 1/L_{\max}$

IFD sampler

Proposal: { fill random empty dyad if $L(x)=L_{obs}$; delete a random tie if $L(x)=L_{obs}+1$

$$\alpha(x \to x') = \min\left\{1, \exp\left(\sum_{A\neq L}\Theta_A \Delta z_A(x) + \Delta z_L V\right)\right\} \qquad V = \left(\Theta_L + \log\left(\frac{L_{\max} - L_{obs}}{L_{obs}+1}\right)\right)$$

Θ_L **value is found from Equilibrium Expectation**

$$P\left(L(x) = L_{obs} \to L(x') = L_{obs}+1\right) = P\left(L(x) = L_{obs}+1 \to L(x') = L_{obs}\right)$$

Byshkin M, Stivala A, Mira A, Krause R, Robins G, Lomi A, J. Stat. Phys. 165: 740-754 (2016)

Existing approaches for MLE

1) Bayesian
2) Geyer-Thompson MCMC MLE
3) Method of Moments (Stochastic approximation)

- Iteratively modify Θ_A
- At many different Θ_A values perform MCMC simulations to draw equilibrium configurations $x(\Theta_A)$

MLE is computationally expensive

Geyer CJ & Thompson EA Journal of the Royal Statistical Society. Series B (Methodological), 657-699 (1992)
Snijders, Tom AB. Journal of Social Structure, 1-40 (2002)

New approach for MLE

If x_{obs} is drawn from $\pi(x, \Theta_A^*)$ (simulated network) then we can estimate true Θ^* from

"Equilibrium Expectation" (EE): $\sum_{x'} P(x_{obs} \to x', \hat{\Theta}_A)\left(z_A(x') - z_A(x_{obs})\right) = 0$

We could prove that

If x_{obs} is large $\Theta^*_A = \hat{\Theta}_A$ If x_{obs} is not large $\Theta^*_A = E_{x_{obs}}(\hat{\Theta}_A)$ (Large sample size)

And obtain very fast MLE!

$x = x_{obs}$ = constant => no MCMC simulation

How to apply this theory if x_{obs} is not a simulated network?

Generous support from the Swiss National Platform of Advanced Scientific Computing (PASC) is gratefully acknowledged

MCMC Equilibrium Expectation

$$\left.\begin{array}{l}\sum_{x'} P(x \to x', \Theta)\left(z_A(x') - z_A(x)\right) = 0 \\ z_A(x) = z_A(x_{obs})\end{array}\right\} \Rightarrow E_\Theta\left(z_A(x)\right) = z_A(x_{obs})$$

Equilibrium Expectation algorithm

1: Initialization: t=0; $x = x_{obs}$; $dz_A = 0$; $dz_A(t=0) = 0$
2: **for** k=1 to m **do**
3: Propose move $x \to x'$ with probability $q(x \to x')$
4: Calculate Metropolis-Hastings acceptance probability $\alpha(x \to x')$
5: **If** $\alpha(x \to x', \theta(t)) > Unif([0,1])$ **then** $dz_A = dz_A + z_A(x') - z_A(x)$ **and perform this move**: $x = x'$
6: **end for**
7: Update of parameters $\theta_A(t+1) = \theta_A(t) - K_A \cdot \mathrm{sgn}(dz_A)(dz_A)^2$
8: Increment t. Save sequences $dz_A(t) = dz_A$; **If** t<M **then** go to step 2

$$\frac{\left|dz_A(t > tc)\right|}{SD(dz_A(t > tc))} < 0.1$$

Simulated networks

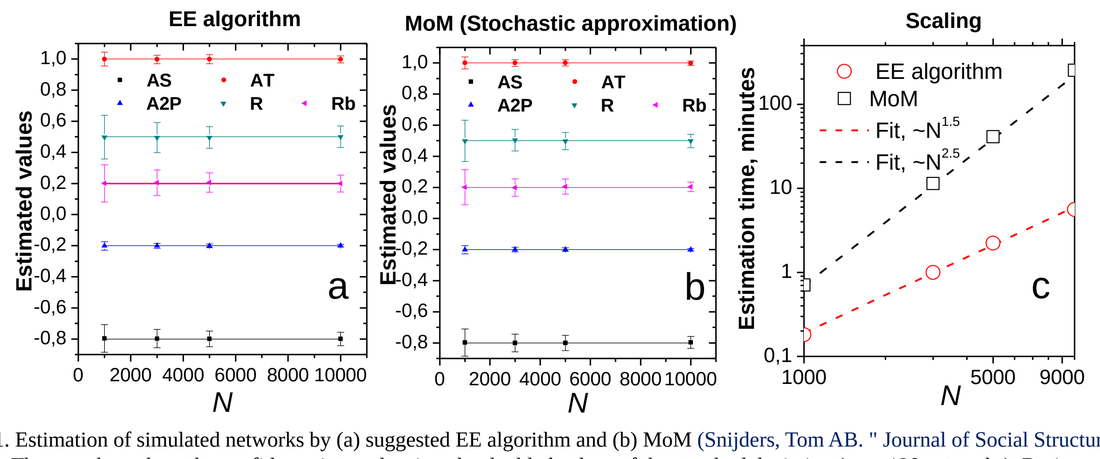

Fig. 1. Estimation of simulated networks by (a) suggested EE algorithm and (b) MoM (Snijders, Tom AB. " Journal of Social Structure (2002): 1-40). The error bars show the confidence intervals, given by doubled values of the standard deviation (over 120 networks). Basic sampler was used. The estimation was performed on Cray XC50 machine available at the Swiss National Supercomputing Centre (CSCS).

AS, AT, A2P, R, Rb are network statistics $z_A(x)$ typical for social networks

Snijders, T.A., Pattison, P.E., Robins, G.L., Handcock, M.S.: New specifications for exponential random graph models. Sociological Methodology **36**(1), 99-153 (2006) ; Robins G, Snijders T, Wang P, Handcock M, & Pattison P (2007) Recent developments in exponential random graph (p*) models for social networks. Social Networks 29(2):192-215.

Biological networks

Arabidopsis Interactome Mapping Consortium. Science, 333(6042):601–607, 2011.
C. von Mering, et al. Comparative assessment of large-scale data sets of protein–protein interactions. Nature, 417(6887):399–403, 2002
T. Rolland et al. A proteome-scale map of the human interactome network. Cell, 159(5):1212–1226, 2014.
X.-T. Huang, et al An integrative C. elegans protein–protein interaction network with reliability assessment based on a probabilistic graphical model. Molecular BioSystems, 12(1):85–92, 2016.
S.-y. Takemura, et al. A visual motion detection circuit suggested by Drosophila connectomics. Nature, 500(7461):175–181, 2013
S. S. Shen-Orr et al, Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics, 31(1):64–68, 2002.

Method	Network	Average sample size	N_c	Avg. estim. time (m)	Elapsed time
EE (IFD sampler)	*A. thaliana* PPI	2160	20	1.1	01 m 50 s
EE (IFD sampler)	Yeast PPI	2617	20	6.6	09 m 07 s
EE (IFD sampler)	Human PPI	4303	20	7.6	10 m 49 s
EE (IFD sampler)	*C. elegans* PPI	5038	20	6.8	09 m 35 s
EE (IFD sampler)	*E. coli* regulatory	418	20	0.6	00 m 43 s
EE (IFD sampler)	*Drosophila* optic medulla	1781	20	4.3	06 m 22 s
SA (IFD sampler)	*A. thaliana* PPI	2160	20	9.2	0 h 34 m 02 s
SA (IFD sampler)	Human PPI	4303	20	49.2	2 h 46 m 54 s
SA (IFD sampler)	Yeast PPI	2617	20	45.6	2 h 02 m 38 s
SA (IFD sampler)	*C. elegans* PPI	5038	20	766.5	25 h 07 m 44 s
SA (IFD sampler)	*E. coli* regulatory	418	20	0.0	0 h 00 m 06 s
SA (IFD sampler)	*Drosophila* optic medulla	1781	20	824.6	72 h 40 m 00 s
SA (basic sampler)	*A. thaliana* PPI	2160	0	—	(time limit)
SA (basic sampler)	Yeast PPI	2617	0	—	(time limit)
SA (basic sampler)	Human PPI	4303	0	—	(time limit)
SA (basic sampler)	*C. elegans* PPI	5038	3	204.5	7 h 40 m 20 s
SA (basic sampler)	*E. coli* regulatory	418	20	1.1	0 h 04 m 06 s
SA (basic sampler)	*Drosophila* optic medulla	1781	0	—	(time limit)
Snowball sampling	*A. thaliana* PPI	490.6	19	26.3	2 h 08 m 24 s
Snowball sampling	Yeast PPI	264.8	19	30.2	3 h 40 m 34 s
Snowball sampling	Human PPI	822.5	18	47.0	3 h 50 m 27 s
Snowball sampling	*C. elegans* PPI	496.4	16	270.7	40 h 00 m 33 s
Snowball sampling	*Drosophila* optic medulla	649.7	15	118.0	7 h 22 m 48 s

Network	Effect	Estimate (95% C.I.) EE (IFD sampler)	SA (IFD sampler)	SA (basic sampler)	Snowball
A. thaliana PPI	AS	2.33 (2.24,2.42)	2.32 (2.23,2.42)	—	2.88 (1.72,3.10)
A. thaliana PPI	AT	1.28 (1.24,1.31)	1.27 (1.23,1.32)	—	0.00 (−0.01,0.01)
A. thaliana PPI	Edge	−14.99 (−15.01,−14.96)	−14.97	—	−14.76 (−16.26,−13.36)
A. thaliana PPI	Isolates	−7.14 (−7.58,−6.69)	−7.12 (−7.58,−6.66)	—	−10.49 (−11.21,−7.95)
Yeast PPI	AS	−0.05 (−0.10,0.01)	−0.05 (−0.10,0.01)	—	0.56 (−0.48,1.29)
Yeast PPI	AT	1.86 (1.81,1.91)	1.86 (1.82,1.90)	—	0.85 (0.18,1.07)
Yeast PPI	Edge	−7.76 (−7.81,−7.71)	−7.76	—	−6.57 (−13.42,−4.88)
Human PPI	AS	1.32 (1.29,1.35)	1.32 (1.29,1.35)	—	1.20 (0.67,2.04)
Human PPI	AT	1.37 (1.35,1.38)	1.37 (1.34,1.39)	—	0.03 (0.02,0.07)
Human PPI	Edge	−11.77 (−11.82,−11.73)	−11.77	—	−9.04 (−13.21,−7.21)
C. elegans PPI	AS	1.04 (1.01,1.07)	1.04 (1.01,1.07)	1.14 (1.03,1.25)	1.06 (0.83,2.16)
C. elegans PPI	AT	1.59 (1.58,1.61)	1.59 (1.57,1.61)	1.52 (1.47,1.57)	0.35 (0.19,0.41)
C. elegans PPI	Edge	−11.03 (−11.08,−10.98)	−10.99	−11.41 (−11.78,−11.04)	−8.82 (−13.40,−7.24)
E. coli regulatory	AS	0.45 (0.32,0.59)	0.44 (0.31,0.57)	0.44 (0.19,0.69)	—
E. coli regulatory	AT	0.78 (0.64,0.93)	0.79 (0.66,0.92)	0.79 (0.61,0.96)	—
E. coli regulatory	Edge	−6.55 (−6.63,−6.47)	−6.53	−6.53 (−7.24,−5.82)	—
Drosophila medulla	AS	0.23 (0.17,0.30)	0.24 (0.18,0.30)	—	1.17 (−0.64,1.58)
Drosophila medulla	AT	1.62 (1.56,1.67)	1.61 (1.57,1.65)	—	1.09 (0.86,1.29)
Drosophila medulla	Edge	−8.14 (−8.19,−8.09)	−8.16	—	−7.70 (−12.51,−5.12)

A. Stivala, et al, Snowball sampling for estimating exponential random graph models for large networks. *Social Networks* 47, 167-188 (2016)

Large social network

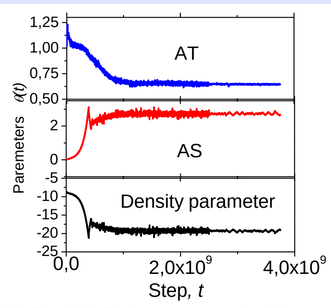

Fig. 2. MLE of ERGM parameters for Livemocha networks with 104103 nodes and 2193083 ties

Zafarani and H. Liu. Social computing data repository at ASU, 2009
Livmocha network datasets, konect.uni-koblenz.de

More presentations

- Complex Networks, March 20-24, 2017, Dubrobnik, Croatia
- Sunbelt INSNA Conference, May 30-June 4, 2017, Beijing, China
- Cambridge Networks Day, 13th June 2017, Cambridge, UK
- PASC17, June 26 - 28, 2017, Lugano, Switzerland
- International Conference on Monte Carlo Methods and Applications, July 3-7, 2017, Montreal, Canada
- International Conference on Computational Social Science, July 10-13, 2017, Cologne, Germany
- Third European Conference on Social Networks, September 26 to 29, 2017, Mainz, Germany